

Wrap-up

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

Why we do gradient descent?

- 1) 실재하지만 알 수 없는 함수 f^* 를 근사하고 싶다.
- 2) 나의 모델(함수) f_θ 의 파라미터 θ 를 조절하자.
- 3) 손실 함수를 최소화 하도록 파라미터 θ 를 조절하자.
- 4) 미분을 통해 gradient를 얻고, loss를 낮추는 방향으로 파라미터를 업데이트

Learning rate

- 실험을 통해 최적화 하는 것이 필요
 - 1) 현재 LR에서 실험이 안정될 경우, 값을 증가: 최적화 속도 증가
 - 2) 현재 LR에서 실험이 불안정(e.g. loss \rightarrow NaN)할 경우, 값을 감소
- 초보자들은 처음에 어떤 값을 정해야 할 지 난감
 - 고민할 바에 그냥 아주 작은 값(e.g. $1e-4$)으로 엄청 오래 돌려도 괜찮다.
- 나중에 Adam Optimizer를 통해 learning rate에 대한 고민을 없앨 수 있음